Categorising Search Sessions: Some Insights from Human Judgments

Tony Russell-Rose
UXLabs
London
UK
+44 (0)7779 936191
tgr@uxlabs.co.uk

Paul Clough
Information School
University of Sheffield
Sheffield, UK
+44 (0)114 222 2664
p.d.clough@sheffield.ac.uk

Elaine G. Toms
Information School
University of Sheffield
Sheffield, UK
+44 (0)114 222 2659
e.toms@sheffield.ac.uk

ABSTRACT
The session is a common unit of interaction that is used in search log analysis. By analysing sessions, it is possible to identify distinct classes of searcher behaviour that can be used to design search applications that better support groups of users based on their expected behaviours. This paper describes an online card sort experiment to investigate how people distinguish between search sessions (i.e., how they group them) to gain insights into their organising principles and to inform the future use of automated approaches, such as clustering. Results show patterns of user behaviour to be the most common way of grouping sessions.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Search process;

General Terms
Performance, Design, Experimentation, Human Factors.

Keywords
Web Search, Information Seeking, Card Sorting, Clustering

1. INTRODUCTION
The session is a common unit of interaction used in the analysis of search logs and for studying patterns of user behaviour [1-3]. A session has a start, that is typically a query, and an endpoint, that is typically a webpage, or arbitrary time-based cut off point, and contains a continuous sequence of user actions with a search engine A session is not equal to a task, as within a single session a user may work on several tasks, and a task may be reflected within multiple sessions [4,5]. It is used for deconstructing tasks, and for optimising search engines. Analysing sessions can show the existence of distinct classes of searcher behaviour [2,6,7] that can be used to design customised search applications based on predicted user behaviour.

The work described in this paper forms part of a wider research project in which we are creating a scheme to categorise search sessions at varying levels of abstraction from low-level actions (e.g., queries and clicks) to higher-level categories of user behaviour (e.g., tactics, tasks and goals). Rather than taking the usual approach of clustering user behavioural patterns [2,6,7], the work reported in this paper takes a user-centered and qualitative approach that involves people grouping 'similar' sessions from a web search engine log using free-sorting or card sorting [8]. This manual analysis was deployed to gain insights into how and why search sessions might be analysed and grouped. The results will be used to inform the use of clustering methods in further work.

The following questions are addressed: (i) How do people distinguish amongst search sessions? (ii) What guiding principles do they use? (iii) What features are commonly used? (iv) Is card sorting a suitable method for eliciting human judgments for categorising sessions? The rest of this paper is structured as follows: Section 2 describes related work; Section 3 describes the research methodology; Section 4 reports initial findings on human performance of the session analysis task; Section 5 discusses the results and Section 6 concludes the paper.

2. RELATED WORK
Various studies have been undertaken to understand how and why people interact with search engines. Such studies have led to the creation of categorisations that describe distinct patterns of use, ranging from individual queries within a session, to entire information seeking episodes. These might reflect patterns of information searching behaviour [9], the types of search tasks that users perform [10], their goals and missions [5], their task switching behaviour [4], or reflect the tasks, needs and goals people are trying to address when using search systems [10,11].

Previous studies have used automated techniques, such as clustering, to identify common user behaviours [2,6,7]. For example, Wolfram et al. [2] selected three separate sources of log data, identified a range of features applicable to each, and then used cluster analysis to identify consistent groups within each data set. These were then manually inspected to identify distinctive characteristics and descriptions of the user behaviours they represented. However, such behavioural patterns are not always consistent across different data sources, and there remains as yet no standardised model of session behaviour. In part this may simply be a reflection of the different contexts in which the data was gathered. But it may also reflect crucial differences in the analysis process. Moreover, the outcome of unsupervised learning approaches, such as clustering, can be highly sensitive to variations in the initial inputs [12].

Recent work has attempted to understand sessions by focusing instead on the direct experience of the searcher. For example, Ye et al. [13] investigated how users understand their own search sessions, using a combination of interviews, manual analysis and

IIiX'14, August 26-30, 2014, Regensburg, Germany.

http://dx.doi.org/10.1145/2637002.2637006

card sorting. Our work complements this by applying a similar qualitative approach, but focuses on exploring the task of session analysis in order to better understand the guiding principles and attributes that people apply in understanding search sessions and what constitutes 'similarity' among them. These insights may help to inform the development of future automated log analysis methodologies.

3. METHODOLOGY

This research extracted a sample of 60 sessions from the Microsoft 2006 Live Search log, and invited a self-selecting sample of ordinary users to group the sessions and label each group using an online card sorting tool. Ideally we would have preferred a face-to-face card sort that provides greater visibility into participant thought processes, but opted instead for a technique that could provide a larger pool of data and increase the likelihood of stable patterns emerging in the output. This procedure was approved by our departmental research ethics review board and was then pilot tested.

Table 1. Sessions in the MSR log by queries and clicks.

Single query	No click	19%
	Single click	36%
	Multiple clicks	7%
Multiple queries	No click	5%
	Single click	8%
	Multiple clicks	25%

Sessions Sample: The log consists of approximately 15 million queries (7,470,915 sessions), sampled over one month from the US Live Search web search engine. Preprocessing of the logs (e.g., sessionisation and removal of IP addresses) has already been carried out. Extracting a simple random sample was deemed inappropriate as the content was skewed toward short, single query interactions. Thus, the log was stratified into six categories (see Table 1) representing varying degrees of user interaction.

Ten sessions from each category were randomly sampled to create an overall pool of 60 sessions (or cards), from which 20 would be selected at random during each card sort. A session was reduced to the key elements: user action (query or click), time stamp (minus the date), keywords with number of results, and clicked URL with its rank. It was also reformatted to be human readable (see Figure 1 for an example).

```
[QUERY] 11:46:17 mercy medical center 10
[QUERY] 11:46:23 mercy medical center ohio 11
[QUERY] 11:46:41 community mercy health partners 15
[CLICK] 11:46:42 http://www.ehealthconnection.com/regions/ 1
```

Figure 1. An example session from the MSR log.

Task: We adapted a card sorting technique, implemented on the web using the OptimalSort (http://www.optimalworkshop.com/) tool. Participants were presented with a set of 20 sessions and then were asked to sort them into groups and label each group. An example set is presented in Figure 2.

The task was divided into two parts. All participants did the control condition, and one of the second (strategy) conditions:

1. A *control condition* consisting of 20 sessions with instructions simply to 'sort them into groups that make sense to you'.

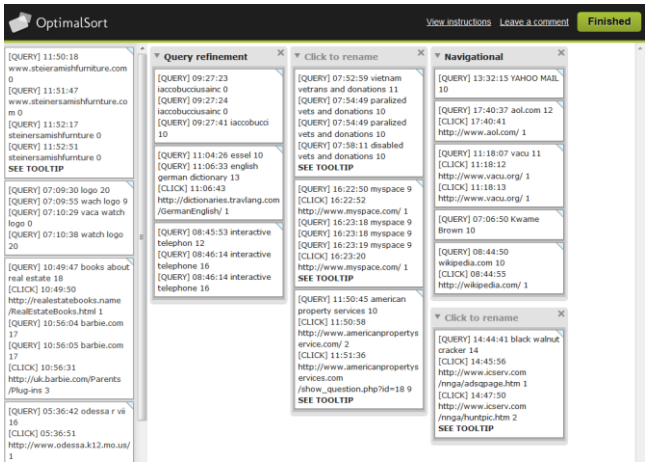

Figure 2. Screenshot of the card sort application.

2. A *strategy condition* consisting of 20 sessions with instructions to sort them according one of the following principles:
   i. By topic (i.e., group sessions with similar themes or subjects together)
   ii. By user behaviour (i.e., group sessions with similar patterns of interaction together)
   iii. By user intent (i.e., group sessions pursuing a similar goal together)

These three strategies were selected as they represent different 'dimensions' of a search session that could provide a basis for grouping.

Participants: Invitations were sent to 174 postgraduate students in the University of Sheffield's iSchool; 43 students responded. Participants were divided into three groups, with each group completing the control condition then one of the three strategy conditions. As an incentive, participants were entered into a prize draw to win a Kindle Fire HD.

Procedure: The system provided an explanation of the task, followed by a brief 'warm up' task so that all participants were familiar with the card sorting tool before beginning the study. They were then given the control condition with its set of 20 sessions, followed by one of the strategy conditions. For each set a random set of sessions was assigned such that no participant received duplicates, and all sessions had an equal chance of being selected. The tool has a drag and drop interface, enabling participants to move each session card from the pool to the desktop and then label. At any time a session could be re-moved and re-labeled. After the session sorting exercise was completed, participants completed a questionnaire to elicit feedback on the strategies and the attributes they found most useful.

4. FINDINGS

Of the 43 participants, 29 completed the initial control condition, and of these 10, 6 and 12 went on to complete each of the topic, structure and goal conditions respectively. The mean (and

median) number of categories per participant, mean number of cards per category and mean time taken are shown in Table 2.

Table 2. Categories, cards and time taken.

Experiment	Mean categories / participant	Mean cards / category	Time taken (range)
Control (N=29)	5.4 (median=5)	3.72	14.43 mins (1.67-38.05)
Topic (N=10)	6.0 (median=5)	3.33	11.6 mins (1.78-21.78)
User behaviour (N=6)	6.5 (median=6.5)	3.08	12.5 mins (5.4-35.83)
User intent (N=12)	4.2 (median=5)	4.80	9.5 mins (3:85-46.15)

4.1 Control Condition

In this condition, the 29 participants sorted the 20 cards into an average of 5.4 groups (Table 2). This resulted in a multiplicity of category (group) names, with minimal apparent overlap between them. However, manually coding them by strategy revealed user behaviour to be the most common approach, followed by topic then user intent (see Table 3, column 2). The coding was performed independently by two judges with 87.2% inter-coder agreement.

Table 3. Strategies used and example group names.

Strategy	% of categories	% of participants	Example category names
Topic	31.4%	15.4%	'transport', 'real estate', 'financial', 'pictures', 'technology' 'health'
User behaviour	37.2%	53.8%	'all query', 'query then click', 'mixed query and click', 'all click'
User intent	17.3%	19.2%	'transactional, 'known item', exploratory'
Other	14.1%	11.5%	'1', 'group1', 'group2'

Table 4. Features used by participants.

Features used	% of participants
keywords	76.9%
number of result pages	26.9%
URLs	50.0%
rank of URLs clicked on	26.9%
timestamps	11.5%
other	11.5%

After the exercise, participants' were asked to describe the strategies they had used to sort the cards. User behaviour was again identified as the most common strategy, followed by user intent then topic (Table 3 column 3). Example category names based on each of these strategies are shown in Table 3 (column 4). Participants were also asked what features they used in completing the task. Keywords were identified as the most commonly used, followed by the content of URLs (Table 4).

4.2 Strategy Conditions

The number of participants completing each of the strategy conditions and the mean number of categories they created is shown in Table 2. The overlap between the category names (where labels suggested by participants were identical or clearly synonymous) is shown in Table 5.

Table 5. Overlap between categories.

Strategy	Overlap
Topic	5/60
Behaviour	1/39
Intent	5/50

A high degree of overlap reflects the degree to which participants share a common approach or mental model. In this instance it is particularly low for the behavioural condition, but this may reflect the lower number of participants and the fact that cards were sampled from a pool of 60 so not all cards were seen by all participants. Participants were also asked to rate the extent to which they found the task a natural way to group sessions (Figure 3) and how difficult they found each task (Figure 4). Grouping by behaviour was slightly preferred in both cases.

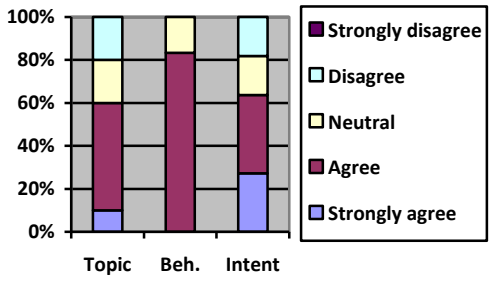

Figure 3. Agreement that <strategy> is a natural way to group the sessions.

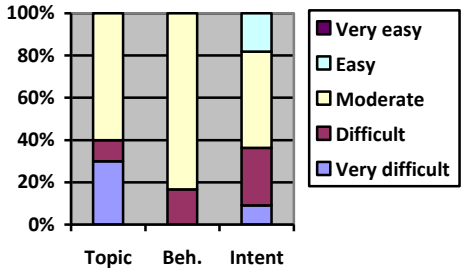

Figure 4. Task difficulty for each strategy.

5. DISCUSSION

5.1 Key Findings

The goal of this study was to better understand better how people distinguish among search sessions and what guiding principles they might apply in grouping similar sessions. This forms an important initial stage in our overall research objective to

investigate the diversity of interaction patterns in search sessions across multiple logs and the development of an appropriate scheme (hierarchical or faceted) to categorise search sessions.

This study is only an initial exploration, and the numbers of participants preclude against definitive conclusions. However, the findings so far indicate that grouping by behaviour is the most common, natural and easiest approach, used by over half the participants in the control condition. In some ways this is understandable, as behaviour is relatively transparent and objective, compared to other, more indirect approaches, such as grouping by topic or intent, which requires some element of inference or subjective judgment (see Fig. 4).

Ironically, grouping by intent was completed more quickly (Table 2), but this may reflect the limited availability of attributes in the data that directly support this approach. The features used by participants may also underline the preference for grouping by user behaviour, since for topic and intent-based approaches it is vital that the content of keywords is analysed and understood, but only 76.9% of participants reported using this feature.

5.2 Methodology

Although the findings provide a unique insight into the task of session analysis, there are various ways in which the methodology could be improved. Firstly, the modifications made to the data to improve readability (by adding annotations for each action) may have biased users towards grouping them by behaviour. Secondly, the minimum number of participants for card sorting is generally considered to be 20-30 [14], so the provision of additional incentives may have been advisable to ensure comparable and sufficient numbers for each of the strategy conditions. This is particularly significant for a task as difficult as session analysis, where there is a multiplicity of ways in which to complete it.

This issue is further reflected in the heterogeneity of the results, with relatively low overlap between the groups created by participants. This may in part reflect the fact that the subject of the analysis is transcripts of 3rd party search sessions, with minimal knowledge of the context involved. The use of a face to face protocol rather than online would have helped facilitate deeper insights into the task itself, particularly the features used, but this may have further compromised participant numbers.

However, these apparent shortcomings draw attention to the wider issue of clustering as a generic approach for search log analysis, in that a given algorithm may reveal ostensibly stable clusters for a given data set but they are by no means the only patterns in that data, and other approaches may reveal entirely different (perhaps contradictory) insights. This raises important questions regarding the repeatability of such studies and validity of their outputs.

6. CONCLUSIONS

This paper describes an experiment to investigate how people distinguish between search sessions and the ways in which they perceive them as 'similar'. This in turn can inform the type of features that should be considered in automated cluster analysis and how the output should be interpreted. We have explored three dimensions of similarity: topic, behaviour, and intent, and found that, although the actual labels assigned to groups varied considerably, the principle of grouping by patterns of user behaviour was the most common, natural and easiest approach. In future work we plan to repeat the experiment with larger numbers of participants.

Acknowledgements

Work partially funded by a Google Research Award project: "Developing a Taxonomy of Search Sessions".

7. REFERENCES

[1] Jansen, B. J. (2006). Search log analysis: What is it; what's been done; how to do it. *Library and Information Science Research*, 28(3): 407–432.

[2] Wolfram, D., Wang, P., and Zhang, J. (2008). Modeling Web session behaviour using cluster analysis: A comparison of three search settings, In *Proceedings of the American Society for Information Science and Technology*, 44(1): 1550-8390.

[3] Weber, I., and Jaimes, A. (2011). Who uses web search for what: and how? In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11)*. ACM, New York, NY, USA, 15-24.

[4] Spink, A., Park, M., Jansen, B.J., and Pedersen, J. (2006). Multitasking during web search sessions. *Information Processing and Management, 42(1): 264-275.*

[5] Jones, R., and Klinkner, K.L. (2008). Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)*. ACM, New York, NY, USA, 699-708.

[6] Chen, H-M., and Cooper, M.D. (2001). Using clustering techniques to detect usage patterns in a web-based information system. *Journal of the American Society for Information Science and Technology*, 52(11): 888–904.

[7] Stenmark, D. (2008). Identifying clusters of user behavior in intranet search engine log files. *Journal of the American Society for Information Science and Technology,* 59(14): 2232-2243.

[8] Coxon, A.P.M (1999). Sorting data: Collection and analysis. Thousand Oaks, CA: Sage Publications

[9] Kellar, M., Watters, C., and Shepherd, M. (2007). A field study characterizing Web-based information seeking tasks. *Journal of the American Society for Information Science and Technology*, 58(7): 999-1018.

[10] Li, Y., and Belkin, N.J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing and Management*, 44(6): 1822-1837.

[11] Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2): 3-10.

[12] Rokach, L., Maimon, O. (2005) Clustering Methods. *The Data Mining and Knowledge Discovery Handbook*, 321-352.

[13] Ye, C., Porcheron, M. & Wilson, M. L. (2013). Studying Extended Session Histories.. In M. L. Wilson, T. Russell-Rose, B. Larsen, P. Hansen & K. Norling (eds.), *EuroHCIR* (p./pp. 23-26), : CEUR-WS.org.

[14] Tullis, T., and Wood, L. (2004), "How Many Users Are Enough for a Card-Sorting Study?" Proceedings UPA'2004 (Minneapolis, MN, June 7-11, 2004).